A COMPARATIVE STUDY ON ATTENTION-BASED RATE ADAPTATION FOR SCALABLE VIDEO CODING

Chia-Ming Tsai[1], *Chia-Wen Lin*[2], *Weisi Lin*[3], *and Wen-Hsiao Peng*[3]

[1]Department of Computer Science, National Chung Cheng University
[2]Department of Electrical Engineering, National Tsing Hua University
cwlin@ee.nthu.edu.tw
[3]School of Computer Engineering, Nanyang Technological University
wslin@ntu.edu.sg
[3]School of Computer Science, National Chiao Tung University
pwen@mail.si2lab.org

ABSTRACT

We conduct subjective tests to evaluate the performance of scalable video coding with different spatial-domain bit-allocation methods, visual attention models, and motion feature extractors in the literature. For spatial-domain bit allocation, we use the selective enhancement and quality layer assignment methods. For characterizing visual attention, we use the motion attention model and perceptual quality significant map. For motion features, we adopt motion vectors from hierarchical B-picture coding and optical flow. Experimental results show that a more accurate visual attention model leads to better perceptual quality. In cooperation with a visual attention model, the selective enhancement method achieves better subjective quality when an ROI has enough bit allocation and its texture is not complex compared to the quality layer assignment. The quality layer assignment method is suitable for region-wise quality enhancement due to its frame-based allocation nature.

Index Terms— Visual attention model, Scalable video coding, Perceptual coding, Video adaptation

1. INTRODUCTION

We live and work in an environment containing heterogeneous networks and devices with various playback abilities. Transporting a video over networks faces many challenges, such as time-varying channel bandwidth, various screen resolutions and resource constraints in playback devices. Thus, adapting video bitstream with optimal quality is very important. The scalable extension of H.264 (a.k.a. SVC) [1] is one of the state-of-the-art video adaptation tools. SVC encodes a video bitstream once and decodes it in various kinds of conditions. In spite of the provision of better visual quality with more enhancement layers, it still cannot ensure that coding improvement matches the human perception. As a result, it might waste many bits in transmitting the data that are not visible to human eyes. To tackle with this problem, it is necessary to combine SVC and perceptual video coding, that is, to transmit the information that human eyes are most interested in to the decoder side first. Thus, it needs to allocate more coding bits to ROIs (regions of interest) of each frame for visual quality improvement. The Selective Enhancement (SE) [2] and Quality Layer Assignment (QLA) [3] methods are major spatial-domain bit allocation methods in SVC.

The SE method improves the visual quality of a region by shifting the coding order of enhancement layers. The most important region will force its enhancement layer to have the highest coding priority so as to be transmitted first. On the contrary, if a region is not as important, its enhancement layer will be coded and transmitted later. However, this method needs to determine the ROIs before applying SE encoding. As a result, if a receiver does not want to receive a perception-based bitstream, it would not be easy to change the ROI settings after video encoding.

The header of an SVC bitstream contains a field '*Priority_Id*' that allows one to define the transmission order of each quality layer. Based on this feature, the QLA method can be implemented in the bitstream extraction process, whereas it is difficult to implement SE in this way. The visual attention map can be regarded as a weighted factor to determine the *Priority_Id* values. The bitstream extraction process then extracts a partial set of video bitstream according to *Priority_Id*. The receiver can decide whether to receive perception-based video bitstream or not and increase the interaction with sender side.

In order to combine SVC with perceptual coding, it is necessary to analyze the video content in pre-encoding time or encoding time. For video content analysis, motion information is a widely used feature in visual attention models [4][5]. Many visual attention models use motion vectors available from motion compensated prediction employed in most video coding standards. The motion attention model (MAM) proposed in [4] considers the motion vector information in the spatial and temporal domains to generate motion attention maps. However in a robust, complete visual attention model, motion is just one of the stimuli features for generating a visual attention map. As for the Perceptual Quality Significant Map (PQSM) [5], in order to get more precise motion information, motion vectors are estimated by an optical flow algorithm [6], making the generated visual attention map more trustable. In SVC, as most video frames are coded as B-frames, the distance of reference frames is different in different temporal levels. As a result, video frames in lower temporal levels might contain many intra MBs due to a long prediction distance. It is therefore interesting to investigate the reliability of motion information obtained from different temporal levels or different motion estimation methods while being adopted in a visual attention model.

In this work, we perform subjective tests to study the visual performance of scalable video coding by combining different spatial-domain bit-allocation methods with different visual attention models and motion features. We only focus on rate adaptation to avoid the influence of temporal and spatial scalabilities, although SVC can support several kinds of scalabilities. We compare two visual attention models: PQSM and MAM. We also integrate the SE and QLA spatial-domain bit-allocation methods to each visual attention model. Besides, we also investigate the influence of motion information on the performance of visual attention model.

The rest of this paper is organized as follows. Section 2 describes the visual attention models and the attention-based rate allocation schemes used in this work. Our methodology of subjective tests is presented in Section 3. Section 4 reports the subjective experimental results and discussions. Conclusions are drawn in Section 5.

2. ATTENTION-BASED RATE ADAPTATION

2.1. Visual attention models

Visual attention models are used to model how a certain image region appeals to human eyes. They are based on top-down (e.g., skin color, face detection, and object appearance) and/or bottom-up (e.g., color contrast, texture contrast, and motion) features. In this work, we evaluate two models: MAM [4] and PQSM [5].

2.1.1. Motion Attention Model

MAM [4] consists of three inductors: intensity inductor, spatial coherence inductor, and temporal coherence inductor. The intensity inductor, *I*, indicates the motion intensity of each macroblock (MB). The spatial coherence inductor, *Cs*, indicates the spatial relationship between neighboring MBs in a video frame. The temporal coherence inductor, *Ct*, indicates the temporal motion vector relationship between adjacent frames. After these inductors are generated, they are integrated into a motion saliency map by

$$B = I \cdot Ct \cdot (1 - I \cdot Cs) \tag{1}$$

In [4], the motion vectors (MVs) obtained from hierarchical B pictures coding (HieBPic) as the motion feature of MAM. For B-frames, there are two motion vectors for each block along the two opposite directions; if we denote them as MV_1 and MV_2, the motion vector is $MV = (|MV_1| + |MV_2|)/2$. In P-frames, a block only has a single MV. For intra MBs in P and B frames, the corresponding motion attention value is set as the maximum value of the motion attention map. Because an I-frame does not contain any MVs, the motion attention map of the previous frame is used when coding the enhancement layer of I frames. Figs. 1(b) and (c) show the motion salience map of MAM for the 2nd frame of *Bus*. In addition to adopting the MVs of HieBPic, we also compare the result with motion information estimated by the optical flow method proposed in [6]. In order to consider the effect of camera motion for a better comparison, the MVs used in the motion attention model have been compensated by global motion compensation using the global motion estimation proposed in [8]. In Fig. 1, "MAM_HieBPic" and "MAM_OF" indicate that the visual attention model used is MAM with motion information obtained from HieBPic and estimated by optical flow, respectively. Similarly, "PQSM_HieBPic" and "PQSM_OF" indicate that the visual attention model used was PQSM with the two motion estimation schemes.

2.1.2. Perceptual Quality Significant Map

PQSM [5] uses three generation procedures, including visual attention features integration, post-processing, and motion suppression to generate a visual sensitivity map. It consists of bottom-up (including color contrast, texture contrast, and motion) and top-down features (including skin color and face detection). After extracting these features, it uses a nonlinear combination method to integrate these features. In the procedure of post-processing, it translates image pixel representation into block representation. For motion suppression, PQSM considers the effect of smooth pursuit eye movement. Figs. 1(d) and (e) depict the attention maps of PQSM for the 2nd frame of *Bus* using MVs directly from HieBPic and estimated by optical flow.

2.1. Quality adaptation mechanisms in SVC

2.2.1. Selective enhancement in SVC

Unlike MPEG-4 FGS that uses bit-plane coding, SVC uses cyclic coding to generate the enhancement layers. However, the cyclic coding method does not consider the visual importance of video content while encoding the enhancement layers. For allocating the coding bits of enhancement layers according to visual importance, we implement the SE method proposed in [2]. If a certain ROI has the highest importance value, the ROI is encoded in the first cyclic coding round. We use the *k*-means clustering algorithm [7] to separate the ROI map into six levels, as shown in Figs. 1 (f)~(i), leading to six shifting levels from 0 to 5.

2.2.2. Quality layer assignment in SVC

The QLA method [3] is one of the tools in JSVM referent software. In the SE approach, additional side information is needed to indicate the coding order of enhancement layer, and the generated bitstream is not standard compliant. The QLA method calculates the distortion between an original frame and its reconstructed frame to identify the RD relationship of each quality layer. The method is fully standard compliant, whereas the drawback is that the resource allocation is frame-based rather than block-based, making the granularity of quality adaptation rather coarse. To implement an attention-based QLA, we incorporate a visual attention map into the original distortion function as follows,

$$D_{\mathrm{ROI}} = \sum_j \sum_i D_{i,j} \times \left(MaxMag_{\mathrm{ROI}} / VAM_{i,j} \right), \tag{2}$$

where $D_{i,j}$ represents the reconstruction distortion, $MaxMag_{\mathrm{ROI}}$ denotes the maximum magnitude of the visual attention map, and $VAM_{i,j}$ denotes the value of the visual attention map at pixel (i, j).

3. METHODOLOGY FOR SUBJECTIVE TESTS

We perform subjective tests to compare the visual qualities of different enhancement methods. We asked 10 subjects to join our experiments. They are graduate students from the College of Engineering of National Tsing Hua University. The test environment is set up in a quiet and comfortable room. The model of the display is EZIO S2431W with a resolution of 1920×1200. Five test sequences, *Bus*, *Football*, *Foreman*, *Mobile*, and *Stefan*, were used in the tests. The test sequences were first encoded using different enhancement methods and are then decoded at different bit-rates with the CIF size and a frame rate of 30 fps. The viewing distance is four times of the frame height. All subjects did not have any prior knowledge or hints about the processing done to the sequences. There were two display windows for every test, showing the reconstructed videos that encoded with and without using an enhancement method, respectively. The two sequences were displayed randomly in the two windows, so the subjects did not know which window was displaying which method's result. The subjects were asked to decide which sequence had better visual quality or there was no difference in visual quality.

In order to understand the influence of the received bit-rate on subjective quality, we choose three bit-rate points to represent low, medium, and high bit-rates for each test sequence. After one bit-rate point test, the subjects were asked to make their decision in 5 s, and then proceed with next bit-rate point test. As shown in Fig. 2, a round of test starts with the low-bit-rate version for a test sequence, and then it is decision time; this follow with the sequence's medium-bit-rate version, and again the decision time; afterward the high-bit-rate version is presented followed by the decision time. For each subjective test, we need to compare four enhancement methods. Because there are four test rounds in a subjective test, the subjects took a rest after two rounds of test to avoid eye fatigue. The time of each subjective test of a subject was limited to be less tan than 30 minutes. Before each test, every subject had at least 5 minutes warm up time to understand what they would be doing.

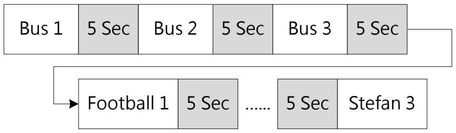

Fig. 2. The order of the play sequence in a test round.

4. EXPERIMENTAL RESULTS

Our experiments include three subjective tests. The first comparison, Exp1, is for SVC coding with and without the SE method. In the second experiment, Exp2, in order to highlight the visual quality of ROI, we removed the enhancement-layer data of regions that have two least important shifting levels, so that more coding bits can be allocated to the ROI using SE. It is also compared with the case without SE. The third experiment, Exp3, compares the results with and without the QLA method when the distortion value is determined by the visual attention map. Since the latest versions of SVC reference software use Medium Grain Scalability (MGS) instead of FGS [9], that are difficult to implement the SE method, we implemented the SE method based on JSVM 8.7 (an FGS version) with the "Palma-CE1-Conditions" setting of coding parameters, and implemented the QLA method based on JSVM 9.15 (an FGS version). For MGS layer coding settings, there were five MGS layers and MGS vector values that were set as 3, 3, 4, 3, and 3, respectively.

Table I shows part of experimental results of three subjective tests, and more results can be found in [10]. The subjective evaluation results show that the performance of PQSM is better than MAM in Exp1 and Exp2. Usually, the SE method degrades the average PSNR quality, as it enhances the quality of specific ROIs while sacrificing the quality of non-ROIs, leading to a globally non-optimal bit allocation. This situation becomes worse when the receiving bit-rate gets higher. In Exp1 and Exp2, when MAM is used, more subjects preferred the videos encoded with HieBPic-based MVs rather than optical-flow-based MVs. This is because MAM_HieBPic allocates coding bits to prediction regions more accurately according to the motion prediction flow of HieBPic compared to MAM_OF (see [10] for subjective comparison). Similarly, in Exp1 and Exp2, the results obtained using PQSM also agree with the MAM case. Compared with the results in Exp1, relatively fewer subjects in Exp2 voted for equal visual quality, because forcing to trade the enhancement layer data of non-ROIs for that of ROIs made most subjects perceive different visual qualities with the two schemes. In Exp1, no matter at which bit-rate, most subjects voted for videos coded without attention-based allocation, whereas in Exp2, more subjects preferred PQSM_OF most of the time except that in *Bus* and *Mobile*. At a medium bit-rate, if an ROI is small and significant, most subjects would vote for the SE method. While at a low bit-rate, the subjective qualities of video coded with and without SE look similar in most cases. However, at a high bit-rate, the SE method seems to degrade the subjective visual quality. The reason is it moves the coding bits from non-ROIs to ROIs, while decoding at high bit-rate points, the quality improvement on an ROI using SE is almost visually not noticeable, whereas the quality degradation on non-ROIs looks relatively visible to most subjects.

Perceptual coding is based on the assumption that trading some coding bits of non-ROIs for enhancing ROIs would make eyes perceive better visual quality. From the experiments, however, this assumption is not always true. In Exp1 and Exp2, although the SE method can improve the subjective quality of main objects in the video, the ROI enhancement is achieved at the cost of degrading non-ROIs. If the quality improvement on the ROIs is not significant enough, the degradation on non-ROIs, rather than the enhancement of ROIs, may in turn dominate the viewing experience of subjects making them feel the video is visually poor. Besides, the accuracy of visual attention map also has significant impact on the performance of an attention-based enhancement scheme. We can conclude that if an ROI in a video frame is visually important and is easy to be enhanced (i.e., improvement can be achieved without consuming many bits), then choosing the SE method is a better choice. Otherwise, using the original SVC coding without attention-based enhancement is enough. Besides, using a more accurate visual attention model or just MAM_HeiBPic will also have good results while combining with the SE method.

Exp3 shows that the PSNR value in each case is the same. Most subjects chose that two videos have equal visual quality. For the few subjects who voted for one bit allocation scheme in this experiment, their decision might just come from their psychological factors. Besides, because the distortion computation in the QLA method is done at frame level, if a video frame has many large visual attention values, it will have a smaller distortion value and its visual quality will become better after enhancement. However this will also degrade the visual quality of other frames which have smaller visual attention values. Therefore, the QLA method is more suitable to combine with event detection method to lay emphasis on the visually important video frames.

5. CONCLUSIONS

Attention-based bit allocation by integrating a visual attention model into the SE or QLA tool of SVC can help transmit human interested data first in video transport, thereby increasing subjective visual quality. The paper presented the initial investigation toward this direction with subjective viewing. The findings provide some useful ground truth, insight and pointers on how to incorporate visual attention models in SVC and possible tradeoffs according to visual content and channel conditions, in the next step of the work. From Exp1 and Exp2, if an ROI in a video frame is visually important and is easy to be enhanced (e.g., objects in video phone and video conference), the SE method is a good choice. Otherwise, attention-based bit allocation may not be able to provide visually significant improvement.

The SE method is suitable in region-wise enhancement in a frame, whereas the QLA method provides enhancement on per-frame basis. Since the SE and QLA methods deal with quality enhancement in different granularities, a combination of the two methods can further enhance the performance of attention-based bit allocation. The accuracy of visual attention model would have a significant influence on the performance of attention-based bit allocation in SVC.

REFERENCES

[1] H. Schwarz, D. Marpe, and T. Wiegand, “Overview of the scalable extension of the H.264/MPEG-4 AVC video coding standard,” *IEEE Trans. Circuits. Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sept. 2007.

[2] W.-H. Peng, T.-H. Chiang, and H.-M. Hang, "Adding selective enhancement in scalable video coding for region-of-interest functionality," in Proc. *IEEE Int. Symp. Circuits Syst.*, May 2006, Greece.

[3] I. Amonou, N. Cammas, S. Kervadec, and S. Pateux, "Optimized rate-distortion extraction with quality layers in the scalable extension of H.264/AVC," *IEEE Trans. Circuits. Syst. Video Technol.*, vol. 17, no. 9, pp.1186–1193, Sept. 2007.

[4] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia*, vol.7, no.5, pp. 907–919, Oct. 2005.

[5] Z. Lu, W. Lin, X. Yang, E. Ong, and S. Yao, "Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation," *IEEE Trans. Image Process.*, vol.14, no. 11, pp.1928–1942, Nov. 2005.

[6] M. J. Black and P. Anandan, “The robust estimation of multiple motions: parametric and piecewise-smooth flow fields,” *Comput. Vis. Image Understand.*, vol. 63, no. 1, pp. 75–104, Jan. 1996.

[7] T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu, "A local search approximation algorithm for k-means clustering," in *Proc. ACM Symp. Comput. Geometry*, 2002, pp.10–18.

[8] C.-T. Hsu and Y.-C. Tsan, "Mosaics of video sequences with Moving Objects," *Signal Process.: Image Commun.*, vol. 19, no. 1, pp. 81-98, Jan. 2004.

[9] H. Kirchhoffer, H. Schwarz, and T. Wiegand, *CE1: Simplified FGS*, Joint Video Team (JVT) of ISO-IEC MPEG & ITU-T VCEG, VTW090 Apr. 2007.

[10] http://www.cs.ccu.edu.tw/~tsaicm/icip2009.htm

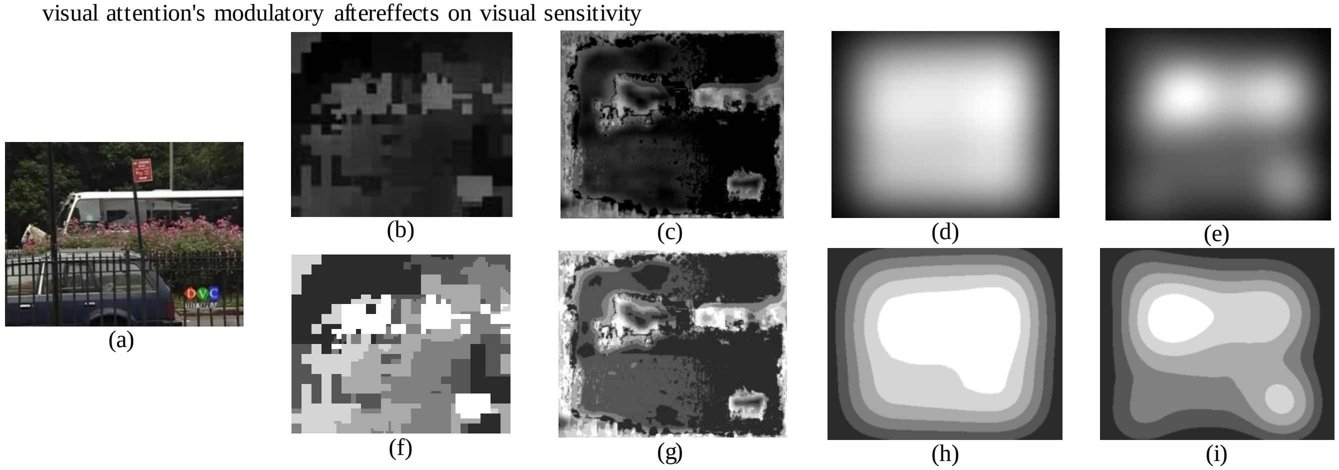

Fig. 1. The 2nd frame of *Bus*: (a) Original image; (b) Saliency map of MAM_HieBPic; (c) Saliency map of MAM_OF; (d) Saliency map of PQSM_HieBPic; (e) Saliency map of PQSM_OF; (f) *k*-means clustering of (b); (g) *k*-means clustering of (c); (h) *k*-means clustering of (d); (i) *k*-means clustering of (e) (*k* = 6). (Note: a region with higher brightness in (b) to (i) means that it is more important to human eyes)

TABLE I The preference opinion scores when using different quality enhancement methods for each subjective test

Test Name	**Exp1**						**Exp2**						**Exp3**					
Sequence Name	**Bus**			**Stefan**			**Bus**			**Stefan**			**Bus**			**Stefan**		
Bit-rate (Kbps)	**300**	**800**	**1500**	**300**	**800**	**1500**	**300**	**700**	**1300**	**300**	**800**	**1500**	**800**	**1400**	**2000**	**700**	**1300**	**1900**
$PSNR_{Ori}$	30.6	29.2	29.2	29.26	33.72	36.57	30.6	31.7	35.7	29.26	31.98	35.26	33.4	34.4	37.5	32.70	34.59	36.42
$PSNR_{MAM\_HieBPic}$	30.5	29.1	29.1	29.19	33.45	35.40	30.5	31.0	33.3	29.28	30.42	32.07	33.7	34.1	36.9	32.33	34.69	36.68
$Q_{MAM\_HieBPic}>Q_{Ori}$	2	1	1	1	3	0	1	3	1	1	1	0	1	3	0	3	1	1
$Q_{MAM\_HieBPic}=Q_{Ori}$	4	7	7	7	5	6	3	6	5	5	1	0	6	6	10	6	8	9
$Q_{MAM\_HieBPic}<Q_{Ori}$	4	2	2	2	2	4	6	1	4	4	8	10	3	1	0	1	1	0
$PSNR_{MAM\_OF}$	30.5	29.1	29.1	29.19	33.48	35.43	30.5	31.0	33.1	29.30	30.41	31.92	33.7	34.1	36.9	32.33	34.69	36.68
$Q_{MAM\_OF}>Q_{Ori}$	0	2	2	2	1	1	2	2	0	2	2	0	3	4	1	2	2	0
$Q_{MAM\_OF}=Q_{Ori}$	5	5	5	5	6	5	5	5	1	5	1	1	6	5	9	8	6	9
$Q_{MAM\_OF}<Q_{Ori}$	5	3	3	3	3	4	3	3	9	3	7	9	1	1	0	0	2	1
$PSNR_{PQSM\_HieBPic}$	30.6	29.1	29.1	29.19	33.49	35.66	30.6	31.2	33.8	29.30	30.91	32.58	33.7	34.1	36.9	32.33	34.69	36.68
$Q_{PQSM\_HieBPic}>Q_{Ori}$	0	0	0	0	2	1	0	1	2	2	3	2	4	1	1	2	1	0
$Q_{PQSM\_HieBPic}=Q_{Ori}$	6	8	8	8	6	5	5	5	5	5	4	5	6	8	7	8	8	10
$Q_{PQSM\_HieBPic}<Q_{Ori}$	4	2	2	2	2	4	5	4	3	3	3	3	0	1	2	0	1	0
$PSNR_{PQSM\_OF}$	30.5	29.1	29.1	29.18	33.53	35.66	30.6	31.0	33.2	29.33	30.61	32.06	33.7	34.1	36.9	32.33	34.69	36.68
$Q_{PQSM\_OF}>Q_{Ori}$	1	2	2	2	1	2	2	3	3	3	5	3	2	3	0	0	1	0
$Q_{PQSM\_OF}=Q_{Ori}$	5	5	5	5	8	7	4	3	4	3	2	6	6	7	10	9	7	10
$Q_{PQSM\_OF}<Q_{Ori}$	4	3	3	3	1	1	4	4	3	4	3	1	2	0	0	1	2	0